ICIC Express Letters
Part B: Applications
Volume **10**, Number **9**, September **2019**

ICIC International ©2019 ISSN 2185-2766
pp. **849–858**

SPEECH RECOGNITION FOR PEOPLE WITH DYSARTHRIA USING CONVOLUTIONAL NEURAL NETWORK

Meisyarah Dwiastuti[1] and Afiahayati[2,*]

[1]Department of Language Science and Technology
Saarland University
Fachrichtung Sprachwissenschaft und Sprachtechnologie, Saarlandes 66123, Germany

[2]Department of Computer Science and Electronics
Universitas Gadjah Mada
FMIPA UGM Sekip Utara, Bulaksumur, Sleman, Yogyakarta 55281, Indonesia
*Corresponding author: afia@ugm.ac.id

Received February 2019; accepted April 2019

Abstract. *Dysarthria is a motoric speech impairment caused by neurological impairment. People with dysarthria often find difficulty in moving their muscles, including the ones around mouth and articulators; thus, the speech produced is not too intelligible. Since speakers with dysarthria are often physically incapacitated, Automatic Speech Recognition (ASR) is more preferred to be implemented in an assistive technology than conventional input method such as switch or keyboard. However, commercial ASRs available today have not reached a good performance when being used by speakers with dysarthria. Convolutional Neural Network (CNN) is well-known for its capability at recognizing pattern, including speech. Its implementation in ASR is able to achieve good performance. In this research, CNN is implemented to build a speaker-dependent isolated-word digit speech recognizer for speakers with dysarthria. The recognizer model is built and evaluated with data of 3 speakers with dysarthria and 1 control speaker. Data speech is provided by UA Speech Database. The best performance obtains average accuracy of 90.43% and NRMSE of 0.1366. Overall, not only speech intelligibility affected the performance, but variety of utterances duration might also have impact on how accurate the classification was.*

Keywords: Convolutional neural network, Dysarthria, Speech recognition

1. **Introduction.** Dysarthria is a neurologic speech disorder characterized by irregularity in articulation of phonemes and amplitude [1]. It may be caused by neuromotor disorder such as celebral palsy, amyotrophic lateral sclerosis, and Parkinson's disease [2]. People's speech with dysarthria is unintelligible for human listeners. Although it may not be the only one, speech intelligibility contributes to measurement of the severity of dysarthria, in which lower intelligibility indicates higher severity [2,3].

Due to neurological condition, people who suffer from dysarthria are usually physically incapacitated. Consequently, conventional assistive technology using switch or keyboard as input is inconvenient for them [4]. Despite their speech impediment, people with dysarthria would rather communicate by speaking than typing because it is less tiring and allows natural communication with eye contact. Therefore, assistive technologies relying on speech as input are introduced, such as STARDUST [5] and VIVOCA [4] and they certainly need to implement reliable speech recognition.

Automatic Speech Recognition (ASR) systems run the process of recognizing spoken language by modeling the relationship between the speech signal and the phone [6]. According to [7], this system may be beneficial for dealing with speech and language impairment. ASR works better than human listening in recognizing consistent articulatory

DOI: 10.24507/icicelb.10.09.849

problem which makes it predictable [2]. In spite of its high intelligibility, the impaired speech may still have characteristics that can be identified by ASR [7]. However, commercial ASRs available today are not designed for people with speech impairment because of different articulation characteristics [3]. Therefore, different approach is needed in order to build an ASR specifically for dysarthric speech.

According to [3], there are several factors that must be considered in order to build dysarthric speech recognizer, namely user's fatigue level, the type of input, the category of ASR technology, and the amount of user and system training provided. Those factors may influence the ASR performance. People with dysarthria are more suitable with the isolated-words recognizer than the long vocabulary one by considering their fatigue level factor. Moreover, a speaker-dependent recognizer shows better performance than others (i.e., speaker-independent, speaker-adaptive) although it causes more training data needed. In speaker-dependent system, each user has to build his own speech profile by providing sufficient amount of utterances as training data and the system only works well for that user [7].

Some previous works have implemented various machine learning methods in building dysarthric speech recognition model, such as Hidden Markov Model (HMM) [4,5,8], Support Vector Machine (SVM) [8], and Multilayer Perceptron (MLP) [9]. Recently, deep learning approach, which is a method of learning in a deep neural network architecture, has shown a great success in pattern recognition, including speech. One of its well-known architecture, Convolutional Neural Network (CNN), has been implemented in speech recognition for people without speech impairment [6,10-13] as a method to build the acoustic model. Meanwhile, for dysarthric speech recognition, CNN architecture has ever been implemented as feature extraction method [14].

For the research described in this paper, CNN is used to build an isolated-word dysarthric speech recognition model. The speech recognition type is speaker-dependent. We use utterances from three subjects with dysarthria varied by their speech intelligibility and one subject without dysarthria. The objective of this research is to build a recognition model with a good performance by investigating the best combination of hyperparameters on CNN architecture and training optimization. This research is an extended version of authors' thesis [15]. In this research, more comprehensive experiments are conducted in order to analyze the results in more detail, especially for CNN-based ASR performance on each subject who has different dysarthria severity and speech intelligibility. This part will be discussed in Section 4 (Experiments and Results) and Section 5 (Discussions).

2. **Methods.**

2.1. **CNN-based speech recognition.** Figure 1 shows the system flowchart of our method to recognize digit. There are two stages, namely training and testing. Training stage is used to train CNN and testing stage is used to evaluate the system.

2.2. **Feature extraction.** In order to train a speech recognition, speech signal must be represented as input features. Before extracting the features, we applied noise reduction to speech files using Audacity. In order to have useful acoustic features, we removed the silence at the beginning and the end of each utterance. Because we wanted all input in the same length, for utterances whose length less than maximum length of our database (i.e., 1870 ms), we added silence at the beginning and at the end of utterance evenly. Therefore, each utterance was exactly at the middle of the speech signal.

Input features have to be extracted and organized as a number of feature maps to be fed into the CNN. In this research, the utterance signal was sampled at 16 kHz and windowed with a 25 ms Hamming window every 10 ms. It resulted 186 frames for each utterance. The 13 MFCC coefficients were then extracted from each frame using 26 filterbanks. These features were represented as a two dimensional feature map in which x

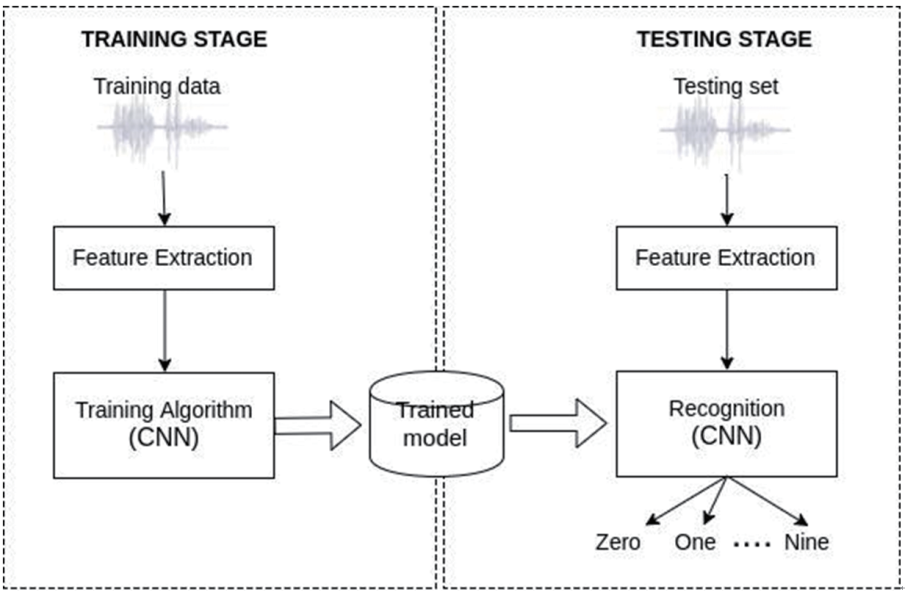


FIGURE 1. Flow of speech recognition method for dysarthric speech using Convolutional Neural Network (CNN)

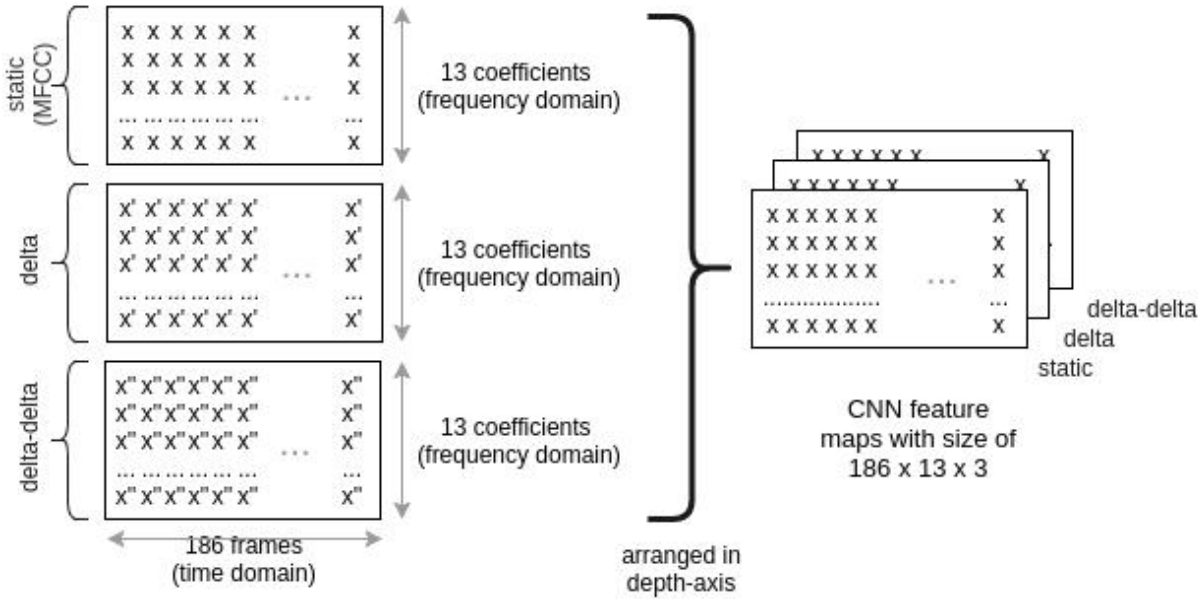


FIGURE 2. CNN feature representation

axis represented time domain (frames) and y axis represented frequency domain (MFCC coefficients). We call this feature map as static feature map. Besides static features, we also used dynamic features to represent temporal changes in an utterance. Therefore, the first and second derivatives of static features were computed and each of both derivative features was also represented as a feature map. Feature map containing first derivative features is called delta feature map and another one containing second derivative features is called delta-delta feature maps. These 3 feature maps were then arranged in the depth axis. Since each feature map had size of 186×13, an utterance was represented as features with size of $186 \times 13 \times 3$. Illustration of the features arrangement is shown in Figure 2.

After features for all utterances had been extracted, they were standardized to scale all the features into the same range. First, mean and standard deviation of features in training data were computed. Then, mean value was subtracted from each feature and the result was divided by standard deviation value. Output of the computation is the

standardized value of the feature. The computation was done to all features in both training and testing data. Those standardized features were the CNN input.

2.3. **CNN architecture and hyperparameter.** In our approach, the CNN consists of an input layer, a pair of convolution layer and pooling layer, a fully-connected layer, and an output layer. The network uses tanh activation function on convolution and fully-connected layers, and softmax activation function on output layer. Max-pooling operation is used on pooling layer. There are 10 neurons in the output and each of them represents one of the labels (i.e., digit).

The hyperparameters of the network are: kernel size, number of feature maps in convolution layer, pooling size and stride, and number of hidden units in fully-connected layer. For initial configuration, the network used kernel size of 6×6, 10 feature maps, pooling size of 2×2 with stride of 2, and 100 hidden units. The best configuration of those hyperparameters were to be selected based on experimental results.

2.4. **Training of CNN.** In this section, we explain the training conditions of the CNN in detail. We trained the network using training data containing pairs of features from the 1st and 2nd utterances and the label of each digit. The initial values of parameters were set as suggested by [16]. The parameters were trained using stochastic gradient descent with 300 epochs and initial learning rate of 0.001. The loss function minimized in the training was categorical cross-entropy.

2.5. **Evaluation criteria.** There are 2 evaluation parameters used in this research. They are defined as follows.

1) Word accuracy: The correct classification proportion predicted by ASR system. This is obtained by computing number of correct classifications of evaluation data and divided it by size of the evaluation data, or written as Equation (1), in which H denotes the number of correct classifications and N denotes number of evaluation data.

$$W.Acc(\%) = \frac{H}{N} \times 100 \tag{1}$$

2) Normalized Root Mean Square Error (NRMSE): It is used to show how close the system results (predictions) are to the target output. Lower NRMSE indicates the system is more accurate. NRMSE is defined as Equation (2).

$$NRMSE = \frac{RMSE}{\max_{target} - \min_{target}} \tag{2}$$

where $RMSE$ is calculated as Equation (3) in which m is the size of evaluation data, n is the size of vocabulary, $target_i$ and $prediction_i$ are target output and prediction output respectively at the i-th neuron. Values of max target and min target are set as maximum and minimum of activation function of the output layer (i.e., softmax function), which are 1 and 0 respectively.

$$RMSE = \sqrt{\frac{\sum_{j=1}^{m} \sum_{i=1}^{n} (target_i - prediction_i)^2}{m \times n}} \tag{3}$$

For model evaluation, we performed 7-fold Cross Validation (CV) on training data. At the end, there were 7 evaluation results and the model performance was the average of those results. This evaluation was performed to every subject data.

3. **Dataset.** We used speech materials provided by UA Speech Database, a collection of utterances by subject with dysarthria caused by celebral palsy [2]. For the research described in this paper, we only used 10-digit (i.e., digit 0-9) utterances of 3 subjects with dysarthria and 1 control subject. The vocabulary and information of the subjects given by UA Speech Database are shown in Table 1. There are 3 utterances for each word uttered by each subject and each utterance has 7 different speech file recorded by different microphones. Therefore, each subject has 210 utterance files. In this research, the third utterance of each digit was used as testing data while the remaining was used as training data. Thus, each subject had testing data containing 70 utterance files and training data containing 140 files.

Table 1. Information of speakers

ID	Sex	Age	Inteligibility level
F02	Female	30	Low (29%)
M05	Male	21	Moderate (58%)
M09	Male	28	High (86%)
CF02 (control)	Female	–	High (100%)

Speaker-dependent recognizer relies on data from one speaker and is used by that speaker only. It is trained and evaluated by data from the same speaker. We wanted to make sure the effectiveness of our recognition model was good enough for various speakers. Therefore, four different subjects were selected based on their speech intelligibility to perform better evaluation.

4. **Experiments and Results.** In this section, we explain two sets of experiments performed to find the best set of hyperparameters in order to build a CNN-based dysarthria speech recognition with highest performance.

4.1. **Experiments on hyperparameters of architecture.** This set of experiments was carried out to identify the best performing configuration of architecture hyperparameters. The experiments were performed sequentially on following hyperparameters: kernel size, pooling size and stride, number of feature maps in convolution layer and number of hidden units in fully-connected layer.

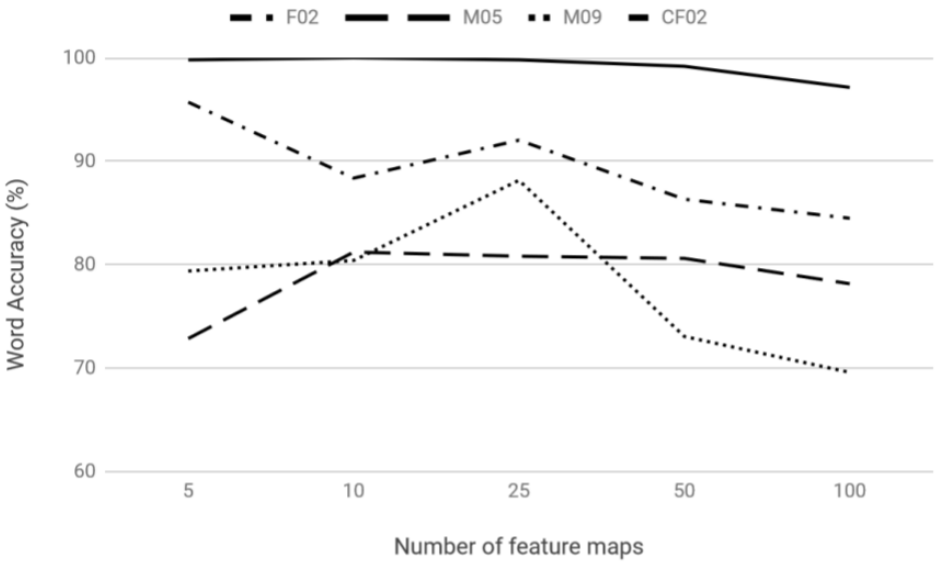
The first experiments (i.e., on kernel size and pooling size) used initial configuration for other hyperparameters. Other experiments (i.e., on number of feature maps and hidden units) set the best performing value resulted from the previous hyperparameter experiments as fixed values for those hyperparameters. The first experiment was to find best value of kernel size. There are two scenarios: when time domain size is greater and when time domain and frequency domain are equal. Kernel size of 10×10 yields the best average accuracy for the first scenario and size of 12×8 yields the best average accuracy for the second scenario. The latter has the best performance with average word accuracy of 83.98%. From the results, we notice that increasing kernel size may rise system accuracy. Moreover, increasing the time domain size also results in better accuracy.

The second experiment was conducted on pooling size and stride. The results in Table 2 show that the system works better when the pooling size is 3×3 regardless of the stride. The highest average accuracy is achieved when pooling size of 3×3 with stride of 1 is used.

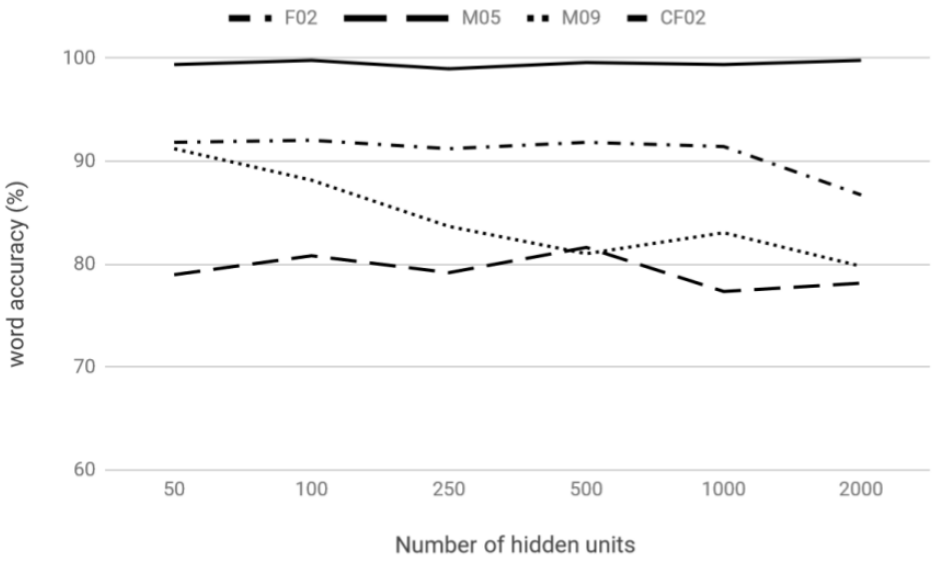
The third experiment was conducted to identify how number of feature maps in convolution layer affects the performance. Figure 3(a) shows the effect of varying number of feature maps on word accuracy for each subject data. Although 25 is not the best value for every subject, the figure shows that using larger number of feature maps drops the

TABLE 2. Results of experiments on pooling size. Other hyperparameters are set as follows: kernel size of 6×6, 10 feature maps, and 10 hidden units.

Size/Stride	Parameter	F02	M05	M09	CF02	Average
$2 \times 2/1$	W.Acc (%)	83.47	75.92	76.73	94.69	82.70
	NRMSE	0.1553	0.1920	0.1967	0.1066	0.1626
$2 \times 2/2$	W.Acc (%)	83.67	72.65	73.47	96.12	81.48
	NRMSE	0.1557	0.1886	0.2021	0.0997	0.1615
$3 \times 3/1$	W.Acc (%)	87.55	76.12	72.65	97.14	83.37
	NRMSE	0.1472	0.1954	0.2095	0.0907	0.1607
$3 \times 3/2$	W.Acc (%)	88.16	77.96	68.57	98.57	83.32
	NRMSE	0.1238	0.1876	0.2078	0.0601	0.1448



(a)



(b)

FIGURE 3. Effects of different number of (a) convolution feature maps and (b) number of hidden units on word accuracy for each subject

accuracy. It shows that the highest average accuracy is achieved when the network uses 25 feature maps.

The last experiment was performed to understand the effect of varying number of hidden units in fully connected layer to the performance. The results show that the best performance is obtained when the fully connected layer has 50 hidden units since it yields the highest average accuracy and lowest NRMSE. Moreover, we can see that the average accuracy declines as the number of units increases, except that 500 units yield slightly better accuracy than 250 units. Meanwhile, the effect of hidden units varies on accuracy of each subject as shown in Figure 3(b). In that figure, the most significant effect is seen on Subject M09 whose accuracy decreases as number of units increases, except that 1000 units yield slightly better accuracy than 500 units. The result of experiments on hyperparameters of architecture was the best performing hyperparameters configuration as follows: kernel size of 12×8, 25 feature maps, pooling size of 3×3 with stride of 1, and 50 hidden units.

4.2. **Experiments on hyperparameters of optimization.** There are two hyperparameters we conducted experiments on, namely learning rate and epoch. The learning rate decay approach used in this research was exponential decay calculated as Equation (4).

$$\eta = \eta_0 k^{\frac{n}{t}} \tag{4}$$

in which η_0 is initial learning rate, k is decay rate, t is decay step, and n is number of parameter updates when decay is applied. We set decay rate as 0.9 and decay step as 1000 parameter updates. Note that the CNN is trained using stochastic gradient descent which updates the parameters for every sample in training data. Since the training set in the CV contains 120 samples, 1000 parameters updates approximately equal to 8 epochs. The result is shown in Table 3. When the learning rate decay is implemented, the average accuracy rises by 1.07% while NRMSE remains the same.

Table 3. Results of experiments on implementation of learning rate decay

Learning rate decay	Parameter	F02	M05	M09	CF02	Average
No	W.Acc (%)	91.84	78.98	91.22	99.39	90.36
	NRMSE	0.1343	0.1823	0.1640	0.0616	0.1355
Yes	W.Acc (%)	94.69	83.27	87.96	99.80	91.43
	NRMSE	0.1259	0.1793	0.1753	0.0660	0.1355

The experiment in epoch was conducted with implementation of learning rate decay because the performance gained was better than when decay was not implemented. We performed experiment with 500 epochs to identify the effect of increasing number of epoch on system performance. From the result in Table 4, we can see that average accuracy slightly declines by 0.05%. Meanwhile, NRMSE decreases quite well not only on the average value, but also on each subject. However, since we were concerned more about word accuracy, we selected 300 as best value for number of epochs.

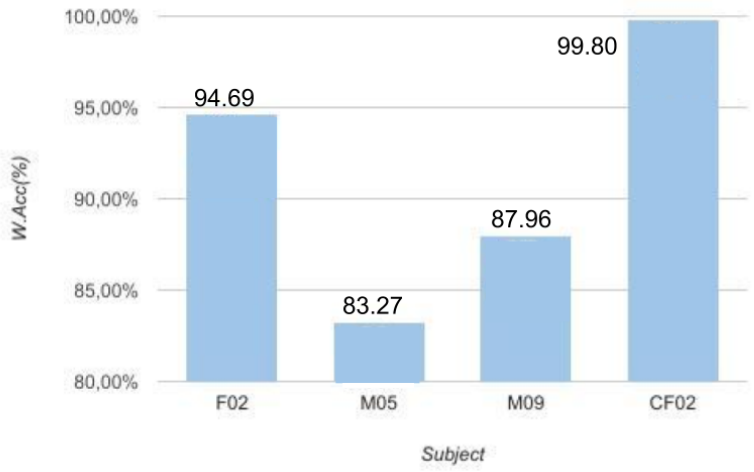
Table 4. Results of experiments on number of epochs

#epoch	Parameter	F02	M05	M09	CF02	Average
300	W.Acc (%)	94.69	83.27	87.96	99.80	91.43
	NRMSE	0.1259	0.1793	0.1753	0.0660	0.1355
500	W.Acc (%)	96.73	80.82	88.16	99.80	91.38
	NRMSE	0.1177	0.1746	0.1651	0.0549	0.1281

5. **Discussions.** In this section, we discuss the CNN-based ASR performance on each subject who has different dysarthria speech intelligibility. Based on experimental results explained, we obtained the best hyperparameters configuration as follows: architecture with convolution kernel size of 12×8, 25 feature maps, pooling size of 3×3 with stride of 1, and 50 hidden units; trained with 300 training epochs, initial learning rate of 0.001, and applying exponential learning rate decay with decay rate of 0.9 for every 1000 parameter updates. The average accuracy gained is 90.43% and average NRMSE is 0.1366.

Figure 4 compares the word accuracy and NRMSE respectively, for each of evaluation dataset (i.e., subject). The system obtained the highest accuracy when being trained and evaluated with the dataset of subject CF02 (control subject), followed with F02, M09, and M05 sequentially. The rank stays the same for NRMSE when the value is sorted in ascending order. It means that, for all subject datasets, when the accuracy is high, NRMSE falls.

The highest accuracy is obtained by subject CF02 whose speech intelligibility is also the highest. Although subjects M05 and M09 are more intelligible than F02, the accuracy of their recognizer is less than 90%. It means that the CNN-based speaker-dependent ASR system performance is not only affected by speech intelligibility of the speaker.



(a)

(b)

FIGURE 4. (a) Word accuracy comparison among subjects; (b) NRMSE comparison among subjects

One factor that can affect the performance is the variation of word utterance duration. For subject F02 the duration varies a lot among different word utterances, but shows less variation in the same word utterances. For instance, the duration of "seven" is around 1.5 second while around 0.8 second for "four". Meanwhile, the duration to utter "seven" varies in range of 1.2-1.5 seconds, and never reaches 1 second. This duration variety is less likely to appear as the speech intelligibility of the speaker is higher. Despite the fact that this factor may have impact on system performance, acoustic features of the utterance have more contribution to how the system can correctly classify the spoken words. It is shown by the accuracy of control subject recognizer which is the highest among all recognizers in spite of having the least variation of word utterances duration on its dataset.

6. **Conclusions.** This paper has studied speaker-dependent isolated digit recognition for people with dysarthria using CNN. We used utterances from subjects with various severity level of dysarthria and one subject without speech impairment to train and evaluate the system. We explored multiple aspects of CNN, namely hyperparameters of architecture and training optimization. Tuning each hyperparameter had different effects on recognition performance for each subject, yet we used the average of all subject performances. The best performing CNN hyperparameters configuration performed the best on control subject dataset and gained average accuracy of 90.43% and average NRMSE is 0.1366. The experimental results showed that not only speech intelligibility affected the performance, but variety of utterances duration might also have impact on how accurate the classification was. In this research, we use CNN as a classifier, in order to gain higher accuracy, a strategy of learning and classifier methods should be attempted, and an example is a hybrid architecture between CNN and SVM or other deep learning algorithms.

REFERENCES

[1] J. Duffy, *Motor Speech Disorders-E-Book: Substrates, Differential Diagnosis, and Management*, Elsevier Health Sciences, https://books.google.co.id/books?id=ATARAAAAQBAJP, 2013.

[2] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. Huang, K. Watkin and S. Frame, Dysarthric speech database for universal access research, *Proc. of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp.1741-1744, 2008.

[3] V. Young and A. Mihailidis, Difficulties in automatic speech recognition of dysarthric speakers and implications for speech-based applications used by the elderly: A literature review, *Assistive Technology*, vol.22, no.2, pp.99-112, 2010.

[4] M. S. Hawley, S. P. Cunningham, P. D. Green, P. Enderby, R. Palmer, S. Sehgal and P. O. Neill, A voice-input voice-output communication aid for people with severe speech impairment, *IEEE Trans. Neural Systems and Rehabilitation Engineering*, vol.21, no.1, pp.23-31, 2013.

[5] M. S. Hawley, P. Enderby, P. Green, S. Brownsell, A. Hatzis, M. Parker, J. Carmichael, S. Cunningham, P. O'Neill and R. Palmer, STARDUST: Speech training and recognition for dysarthric users of assistive technology, *The 7th European Conference for the Advancement of Assistive Technology (AAATE 2003)*, pp.959-963, 2003.

[6] D. Palaz, M. Magimai-Doss and R. Collobert, Analysis of CNN-based speech recognition system using raw speech as input, *Proc. of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, vol.2015, pp.11-15, 2015.

[7] P. Kitzing, A. Maier and V. L. Åhlander, Automatic Speech Recognition (ASR) and its use as a tool for assessment or therapy of voice, speech, and language disorders, *Logopedics, Phoniatrics, Vocology*, vol.34, no.2, pp.91-96, 2009.

[8] M. Hasegawa-Johnson, J. Gunderson, A. Penman and T. Huang, HMM-based and SVM-based recognition of the speech of talkers with spastic dysarthria, *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol.3, pp.1-4, 2006.

[9] S. R. Shahamiri and S. S. B. Salim, Artificial neural networks as speech recognizers for dysarthric speech: Identifying the best-performing set of MFCC parameters and studying a speaker-independent approach, *Advanced Engineering Informatics*, vol.28, no.1, pp.102-110, 2014.

[10] O. Abdel-Hamid, A. R. Mohamed, H. Jiang and G. Penn, Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition, *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp.4277-4280, 2012.

[11] C. K. Dewa and Afiahayati, Suitable CNN weight initialization and activation function for javanese vowels classification, *Procedia Computer Science*, vol.144, pp.124-132, 2018.

[12] R. Adam, C. K. Dewa and Afiahayati, Recognizing arabic letter utterance using convolutional neural network, *Proc. of the 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, pp.181-186, 2017.

[13] C. K. Dewa, Javanese vowels sound classification with convolutional neural network, *Proc. of International Seminar on Intelligent Technology and Its Applications (ISITIA)*, 2017.

[14] T. Nakashika, T. Yoshioka, T. Takiguchi, Y. Ariki, S. Duffner and C. Garcia, Dysarthric speech recognition using a convolutive bottleneck network, *IEEE International Conference on Signal Processing (ICSP)*, pp.505-509, 2014.

[15] M. Dwiastuti, *Pengenalan Ucapan Pada Orang Dengan Dysarthria Menggunakan Convolutional Neural Network (in English: Speech Recognition for People with Dysarthria Using Convolutional Neural Network)*, Bachelor Thesis, Universitas Gadjah Mada, Indonesia, 2017.

[16] X. Glorot and Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, *Proc. of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS'10)*, 2010.